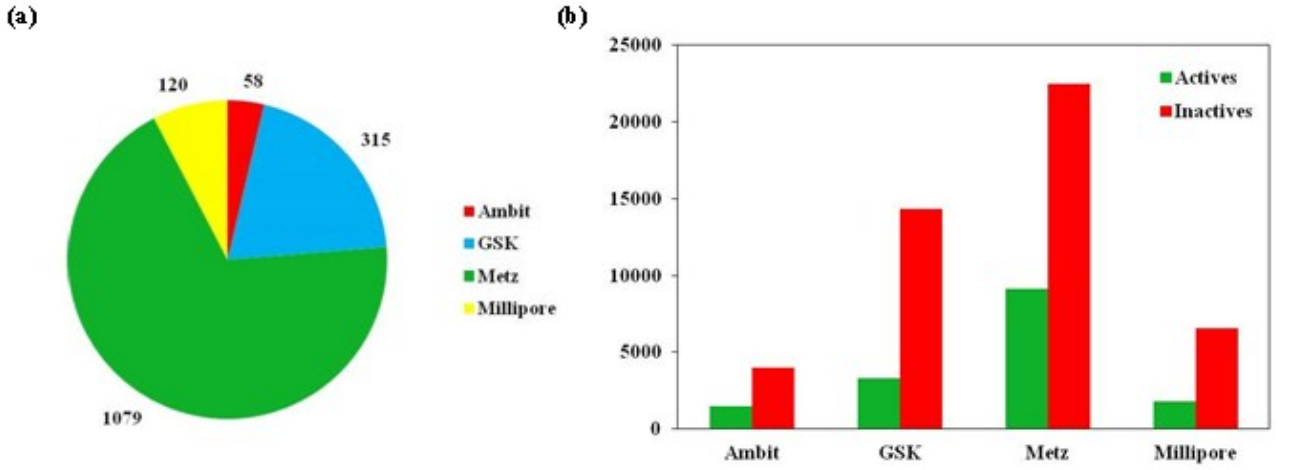

Figure S1. Distribution of the dataset **(a)** Pie chart representing the distribution of the ligands in different datasets **(b)** Bar plots showing the distribution of actives and inactives in each dataset

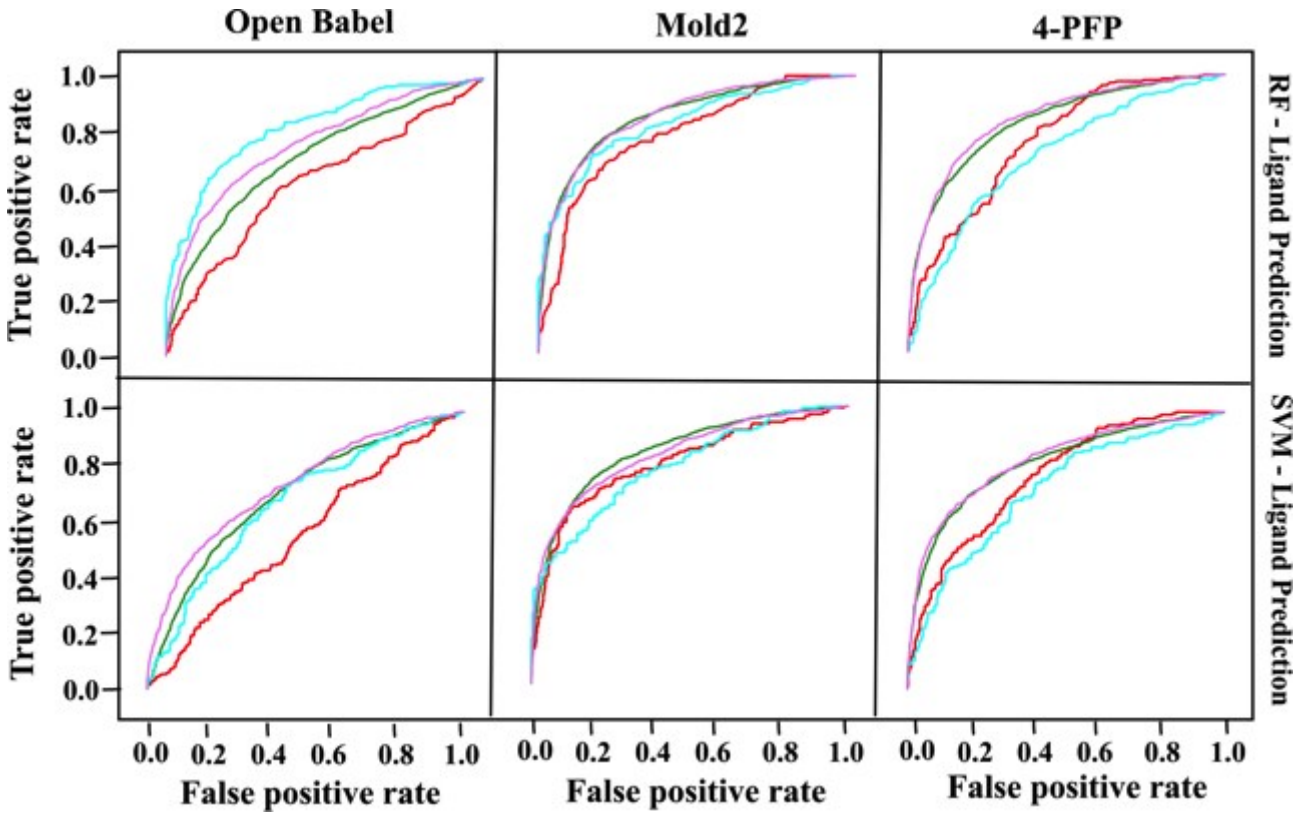

Fig. S2 ROC curves showing the performances of proteochemometric (PCM) models based on different datasets, machine learning algorithms (RF – Random Forests and SVM – Support Vector Machines) and descriptors. Red, forest green, cyan and violet coloured curves correspond to the performances of Ambit, Metz, Millipore and GSK datasets.

Table S1. Experimental kinase inhibition data extracted from different sources

Source	Interaction data	Inhibitors	Kinases	Data points
Karaman et al.[a]	Dissociation constant [K_d]	38	317	12046
Metz et al.[a]	Negative log of Inhibition constant [pK_i]	1497	172	107791
Davis et al.[a]	Dissociation constant [K_d]	72	442	31824
Kinase SARfari[a]	Dissociation constant [K_d]	2453	359	24351
Chembl - GSK[b]	Inhibition %	367	224	82208
Chembl - Millipore[a]	Residual activity (%)	158	234	36972

[a] Experimental measurements at 10 μM concentration
[b] Experimental measurements at 1 μM concentration

Table S2. Sigma and cost parameters used in training SVM models based on different ligand descriptors

Descriptor	Sigma	Cost
Ligand prediction set		
Open Babel	0.0156250	20
Mold2	0.00390625	60
4 point pharm	0.00390625	40
Target prediction set		
Open Babel	0.0156250	30
Mold2	0.00390625	70
4 point pharm	0.00390625	70

Table S3. Performance of proteochemometric (PCM) models in predicting activities of ligand test set. Sensitivity, specificity, accuracy percentage and Matthews coefficients resulting from PCM models based on different ligand descriptors and different machine learning approaches

Ligand descriptors	Method	Sensitivity	Specificity	Accuracy	Matthews	Kappa	AUC
Cross-validation							
Open Babel	SVM[a]	0.53	0.91	0.81	0.47	0.47	0.83
	RF[b]	0.41	0.96	0.82	0.48	0.44	0.86
Mold2	SVM[a]	0.62	0.91	0.84	0.56	0.56	0.87
	RF[b]	0.43	0.96	0.83	0.5	0.47	0.88
4-PFP[c]	SVM[a]	0.58	0.92	0.84	0.54	0.54	0.86
	RF[b]	0.43	0.96	0.83	0.49	0.46	0.87

External prediction							
Open Babel	SVM[a]	0.45	0.83	0.74	0.28	0.28	0.70
	RF[b]	0.18	0.97	0.78	0.25	0.20	0.73
Mold2	SVM[a]	0.60	0.90	0.83	0.52	0.51	0.83
	RF[b]	0.37	0.97	0.83	0.47	0.42	0.85
4-PFP[c]	SVM[a]	0.52	0.91	0.82	0.47	0.47	0.82
	RF[b]	0.32	0.97	0.81	0.42	0.36	0.83

[a]Support Vector Machines [b]Random Forests [c]4 Point Pharmacophoric Fingerprints

Table S4. Performance of proteochemometric (PCM) models in predicting activities of target test set. Sensitivity, specificity, accuracy percentage and Matthews coefficients resulting from PCM models based on different ligand descriptors and different machine learning approaches.

Ligand descriptors	Method	Sensitivity	Specificity	Accuracy	Matthews	Kappa	AUC
Cross-validation							
Open Babel	SVM[a]	0.55	0.91	0.82	0.49	0.49	0.83
	RF[b]	0.43	0.96	0.83	0.49	0.46	0.86
Mold2	SVM[a]	0.58	0.91	0.83	0.52	0.52	0.85
	RF[b]	0.40	0.97	0.83	0.49	0.45	0.87
4-PFP[c]	SVM[a]	0.59	0.92	0.84	0.54	0.54	0.86
	RF[b]	0.42	0.96	0.83	0.49	0.46	0.87
External prediction							
Open Babel	SVM[a]	0.98	0.01	0.74	-0.01	0	0.48
	RF[b]	0.30	0.97	0.80	0.39	0.34	0.82
Mold2	SVM[a]	0.10	0.90	0.70	0	0	0.51
	RF[b]	0.33	0.97	0.81	0.42	0.37	0.83
4-PFP[c]	SVM[a]	0.11	0.83	0.65	-0.06	-0.06	0.51
	RF[b]	0.33	0.97	0.81	0.41	0.36	0.82

[a]Support Vector Machines [b]Random Forests [c]4 Point Pharmacophoric Fingerprints

Table S5. Predictions of kinase targets resulting from internal cross-validation grouped according to the kinase families

Family	Number of kinases	Actives	Inactives	Sensitivity	Specificity	Accuracy	Matthews	AUC
AGC	5	810	4055	0.23	0.97	0.87	0.32	0.84
CAMK	11	1792	4524	0.45	0.96	0.83	0.5	0.86
CK	2	483	1551	0.36	0.96	0.82	0.43	0.84
CMGC	15	3297	7353	0.39	0.94	0.81	0.4	0.82
OPK	3	81	406	0.2	0.95	0.85	0.34	0.83
STE	11	1093	4194	0.4	0.97	0.83	0.47	0.86
TK	23	4999	15817	0.37	0.96	0.83	0.44	0.87
TKL	5	209	638	0.4	0.94	0.82	0.4	0.83

Table S6. Predictions of kinase targets resulting from external test set validation, grouped according to the kinase families

Family	Number of kinases	Actives	Inactives	Sensitivity	Specificity	Accuracy	Matthews	AUC
CAMK	3	417	1354	0.44	0.96	0.87	0.47	0.86
CKI	1	298	773	0.33	0.97	0.79	0.43	0.80
CMGC	3	142	707	0.41	0.94	0.86	0.42	0.82
STE	5	553	1909	0.37	0.97	0.82	0.43	0.82
TK	7	1426	3935	0.32	0.96	0.77	0.37	0.82
TKL	1	129	242	0.36	0.92	0.73	0.35	0.77

Table S7. Performance of proteochemometric (PCM) models in predicting activities of ligand test sets. AUC values resulting from cross-validation and external prediction are grouped according to the sources from which the ligands are extracted.

Ligand descriptors	Method	Cross-validation				External prediction			
		Ambit	Metz	Millipore	GSK	Ambit	Metz	Millipore	GSK
Open Babel	SVM[a]	0.83	0.82	0.84	0.84	0.55	0.70	0.67	0.73
	RF[b]	0.86	0.85	0.89	0.86	0.62	0.72	0.82	0.76
Mold2	SVM[a]	0.87	0.86	0.87	0.86	0.80	0.84	0.78	0.83
	RF[b]	0.88	0.87	0.89	0.87	0.79	0.85	0.83	0.86
4-PFP[c]	SVM[a]	0.87	0.85	0.87	0.85	0.77	0.81	0.72	0.82
	RF[b]	0.88	0.86	0.89	0.87	0.76	0.83	0.71	0.84

Table S8. Performance of RF models based on different conformations of the ligands SKI-606 (Bosutinib) and TAE-684, whose x-ray structures are known. AUCs and Matthews (in parenthesis) resulting from internal cross-validation of the models and prediction probabilities of the active class are shown. PDB mentioned in the table corresponds to the conformation found in x-ray structure. Lowest, second lowest, medium and highest represent the conformations, ranked according to their energy values.

Ligand conformation used in modelling	AUC (Matthews)	Prediction probabilities of the active class
SKI -606 (Bosutinib)		**Interactions of SKI-606 with ABL1 kinase**
SKI-606 (PDB)	0.87 (0.49)	0.52
SKI-606 (lowest)	0.87 (0.49)	0.58
SKI-606 (second lowest)	0.86(0.48)	0.55
SKI-606 (medium)	0.86 (0.48)	0.50
SKI-606 (highest)	0.86 (0.47)	0.49
TAE-684		**Interactions of TAE-684 with ALK kinase**
TAE-684 (PDB)	0.87 (0.49)	0.82
TAE-684 (lowest)	0.87 (0.49)	0.81
TAE-684 (second lowest)	0.86 (0.48)	0.81
TAE-684 (medium)	0.86 (0.48)	0.83
TAE-684 (highest)	0.86 (0.48)	0.76

Table S9. Prediction probabilities of the active class after excluding the fingerprints used for interpretation.

Excluded pharmacophoric fingerprint	Ligand-kinase combination shown in interpretation	Prediction probabilities of the active class	
		With fingerprint	Without fingerprint
AARR	TAE-684 and ALK	0.81	0.29
AAAR	TAE-684 and ALK	0.81	0.42
AAAR	SKI-606 and ABL1	0.58	0.27
ADRR	SKI-606 and ABL1	0.58	0.47
AARR	Vandetanib and STK10	0.63	0.52
AAAR	Vandetanib and STK10	0.63	0.49
AADR	Vandetanib and STK10	0.63	0.52
ADRR	Vandetanib and STK10	0.63	0.50